April 2005

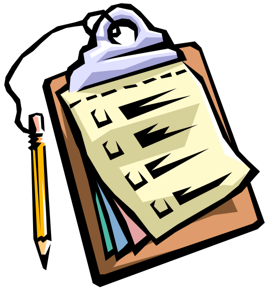
AIX AK April 2005
Forschungszentrum Karlsruhe

IBM Systems Group

IBM Storage with Linux 2005

Alexander Warmuth
ATS EMEA Storage

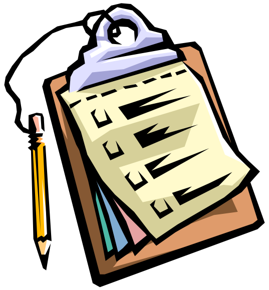
Topics



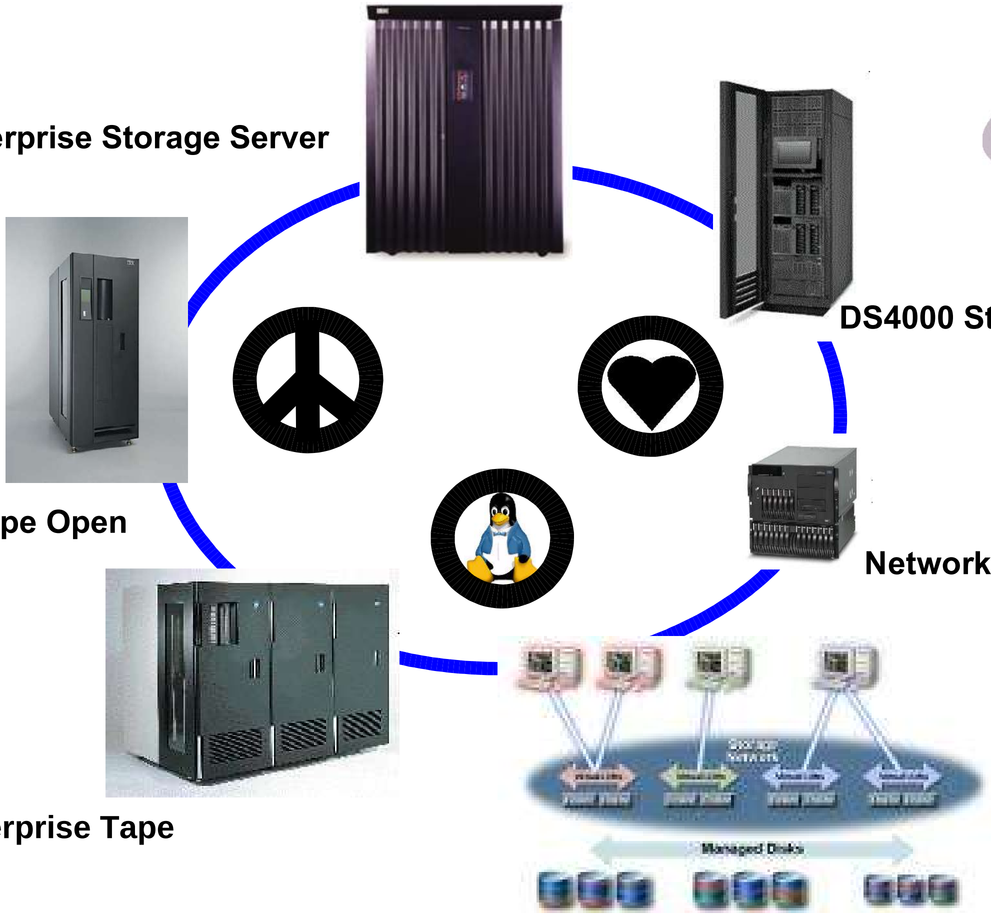
- What is supported
- The Linux SCSI subsystem
- Linux Kernel version 2.6
- Multipathing scenarios
- Tape specifics

What Is Supported

- **What is supported**
- The Linux SCSI subsystem
- Linux Kernel version 2.6
- Multipathing scenarios
- Tape specifics



IBM Storage Support for Linux



Enterprise Storage Server

DS4000 Storage Servers

Linear Tape Open

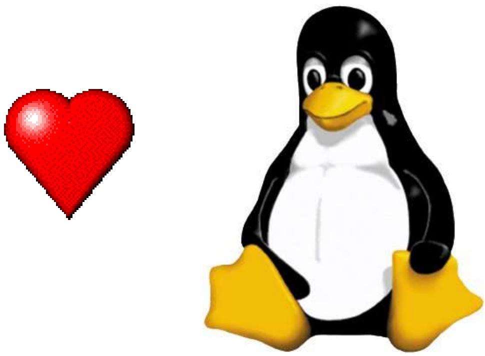
Network Attached Storage

Enterprise Tape

Storage Virtualization

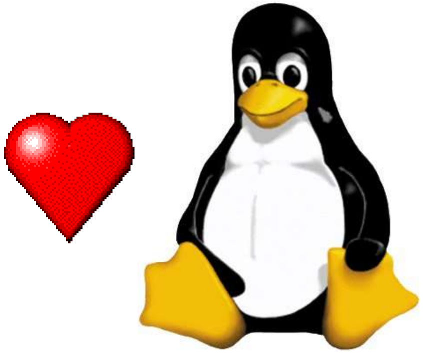
pLinux Support Disk

- ESS: SLES8, SLES9, RH-EL 3
  - SDD available
  - JS20, p5 and OP: SLES9, RH-EL 3
  - Remote boot supported
- DS6000, DS8000: SLES8, SLES9, RH-EL 3
  - SDD available
  - JS20: SLES8, SLES9, RH-EL 3
  - p5 and OP: RH-EL 3 only
  - Remote boot supported
- DS4000: SLES8, SLES9, RH-EL 3
  - Emulex Multipulse driver for multipathing
  - SLES 9 single path only
  - Remote boot with JS20, others require RPQ

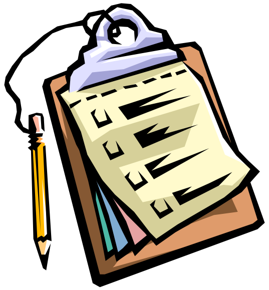


pLinux Support Tape

- **LTO: SLES 8, SLES 9, RH-EL 3**
  - **Data Path (for 3584) and Media Changer failover supported**
- **359x: SLES 8, SLES 9, RH-EL 3**
  - **Data path failover supported for 3592**
- **Parallel SCSI attachment also supported**
- **Advanced**
  - **`IBMtape` device driver**
  - **`IBMtapeutil`**



The Linux SCSI Subsystem



- What is supported
- **The Linux SCSI subsystem**
- Linux Kernel version 2.6
- Multipathing scenarios
- Tape specifics

Linux Device Addressing

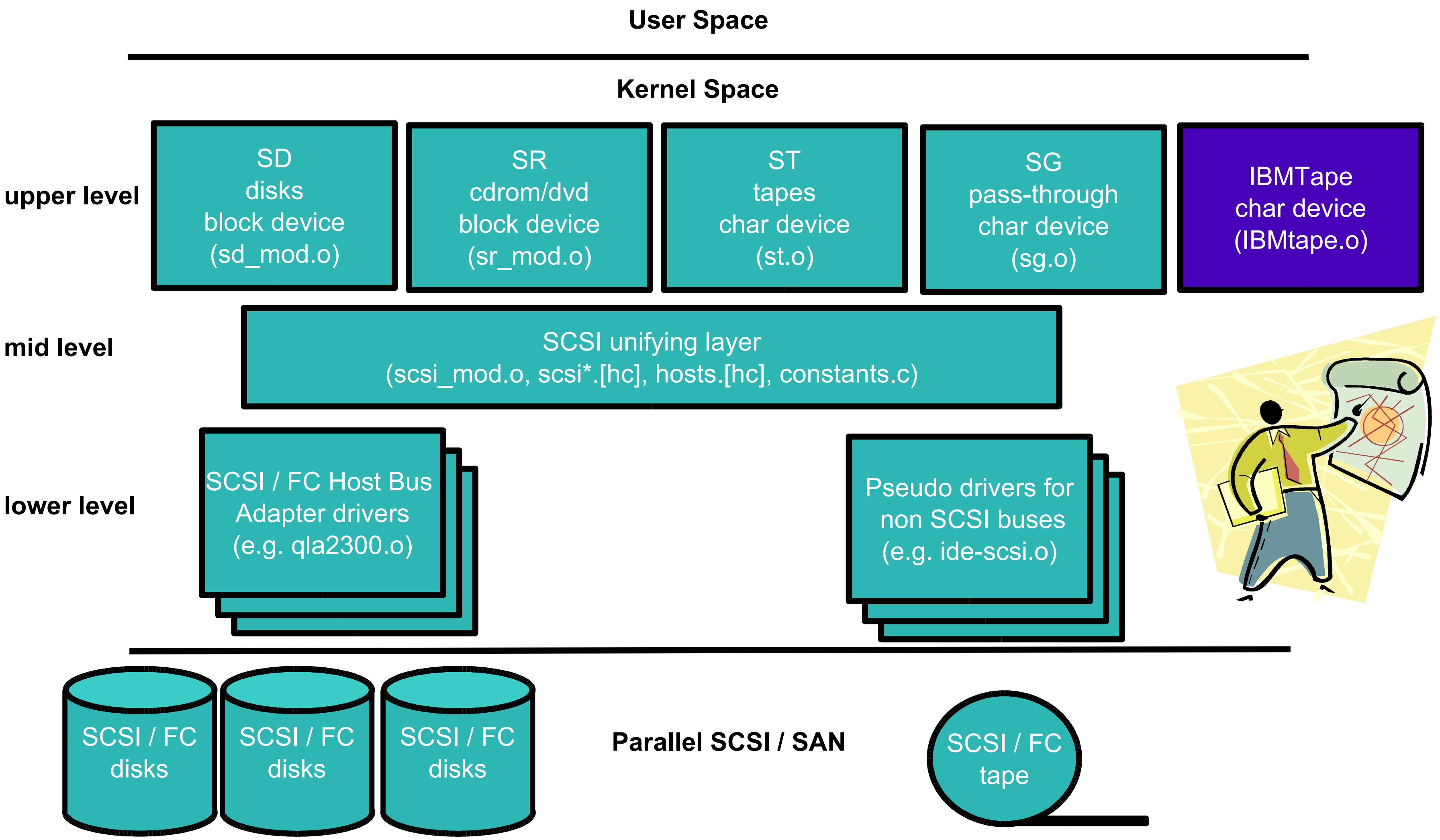
Everything is a file!

```
brw-rw----     1 root      disk         8,    0 2003-03-14 14:07 /dev/sda
brw-rw----     1 root      disk         8,    1 2003-03-14 14:07 /dev/sda1

brw-rw----     1 root      disk         3,    0 2003-03-14 14:07 /dev/hda

crw-rw----     1 root      disk         9,    0 2003-03-14 14:07 /dev/st0
crw-rw----     1 root      disk         9,   96 2003-03-14 14:07 /dev/st0a
crw-rw----     1 root      disk         9,   32 2003-03-14 14:07 /dev/st0l
crw-rw----     1 root      disk         9,   64 2003-03-14 14:07 /dev/st0m
```

Design



User Space

Kernel Space

upper level

SD
disks
block device
(sd_mod.o)

SR
cdrom/dvd
block device
(sr_mod.o)

ST
tapes
char device
(st.o)

SG
pass-through
char device
(sg.o)

IBMTape
char device
(IBMtape.o)

mid level

SCSI unifying layer
(scsi_mod.o, scsi*.[hc], hosts.[hc], constants.c)

lower level

SCSI / FC Host Bus Adapter drivers (e.g. qla2300.o)

Pseudo drivers for non SCSI buses (e.g. ide-scsi.o)

SCSI / FC disks

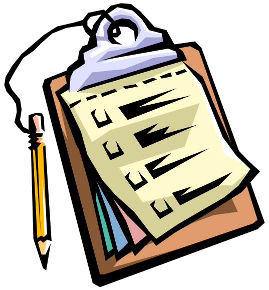
SCSI / FC disks

SCSI / FC disks

Parallel SCSI / SAN

SCSI / FC tape

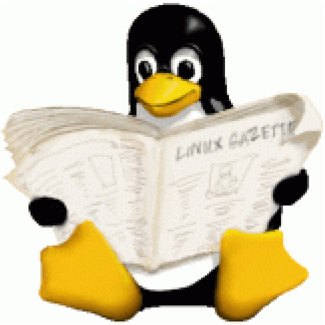
Linux Kernel Version 2.6



- What is supported
- The Linux SCSI subsystem
- Linux Kernel version 2.6
- Multipathing scenarios
- Tape specifics

Storage Changes in Linux Kernel 2.6

- Increased number of SCSI devices
- Persistent device names
- Improved hotplugging
- Native multipathing
- LVM 2
- Improved I/O performance
- Larger devices and filesystems



Wellknown Linux SCSI Limitations

- Limited numbe
  - Up to 256 S
  - Up to 256 SCSI g
  - Up to 32 tape drives
- Gaps in LUN sequen
- Limited “on-the-fl
- Device re-or

Fixed with Kernel 2.6

Still there

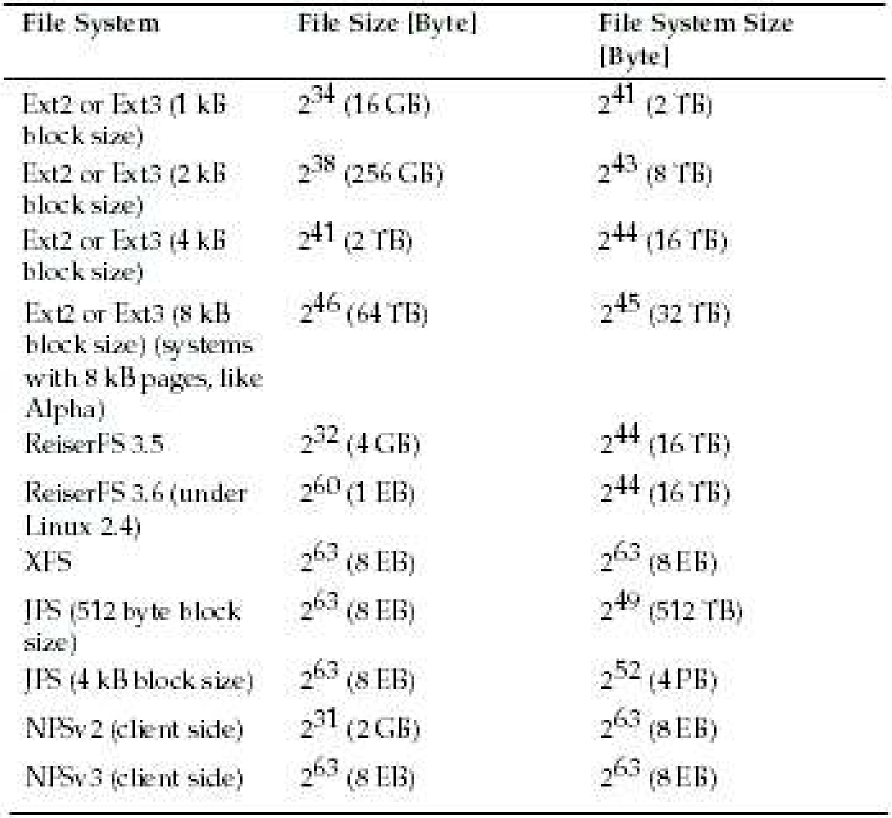
Fixed with Kernel 2.6

Conditionally fixed with Kernel 2.6

Other Problems and Pitfalls

- **Multiple LUN support of RH-EL**
- **DS4000 Specific**
  - **QLogic failover driver configuration**
  - **Potential LUN thrashing**
  - **UTM (Access LUN)**
- **ESS, DS6000, DS8000 Specific**
  - **SDD and LVM, ext3**
  - **Mounting PPRC targets**
  - **DS6000 Preferred Path**

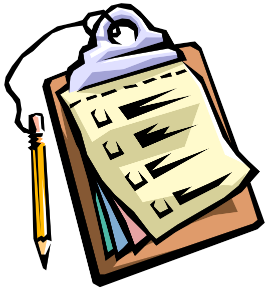
Large filesystems support



File System	File Size [Byte]	File System Size [Byte]
Ext2 or Ext3 (1 kB block size)	2^{34} (16 GB)	2^{41} (2 TB)
Ext2 or Ext3 (2 kB block size)	2^{38} (256 GB)	2^{43} (8 TB)
Ext2 or Ext3 (4 kB block size)	2^{41} (2 TB)	2^{44} (16 TB)
Ext2 or Ext3 (8 kB block size) (systems with 8 kB pages, like Alpha)	2^{46} (64 TB)	2^{45} (32 TB)
ReiserFS 3.5	2^{32} (4 GB)	2^{44} (16 TB)
ReiserFS 3.6 (under Linux 2.4)	2^{60} (1 EB)	2^{44} (16 TB)
XFS	2^{63} (8 EB)	2^{63} (8 EB)
JFS (512 byte block size)	2^{63} (8 EB)	2^{49} (512 TB)
JFS (4 kB block size)	2^{63} (8 EB)	2^{52} (4 PB)
NFSv2 (client side)	2^{31} (2 GB)	2^{63} (8 EB)
NFSv3 (client side)	2^{63} (8 EB)	2^{63} (8 EB)

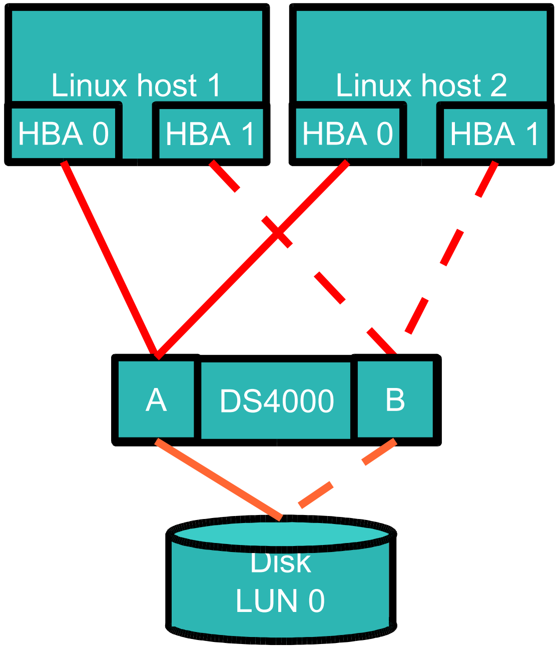
- **Linux Kernel Limits**
  - **Max file size: 2 TB (2^41 bytes)**
  - **Max file system size: 8 ZB (2^73 bytes)**

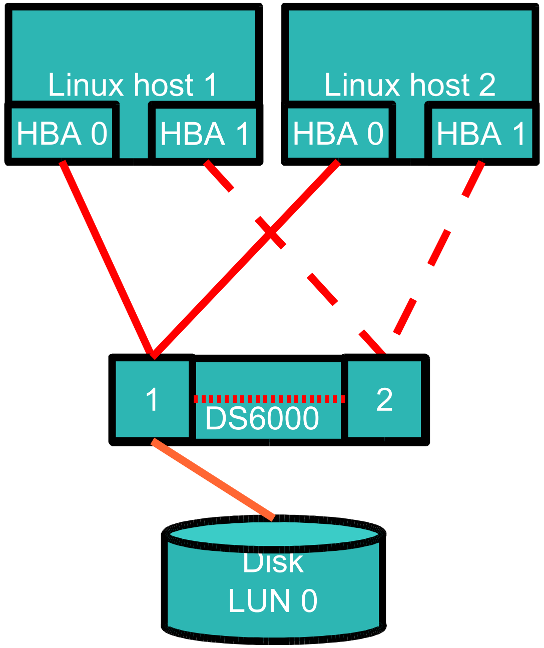
Multipathing Scenarios

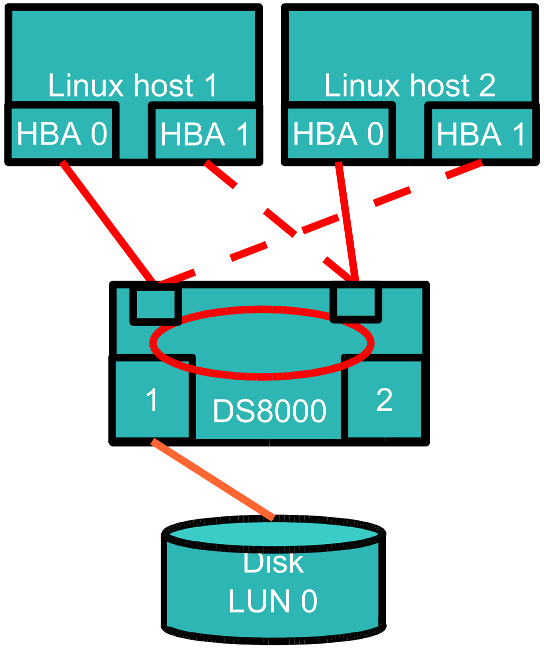


- What is supported
- The Linux SCSI subsystem
- Linux Kernel version 2.6
- **Multipathing scenarios**
- Tape specifics

Multipathing Concepts



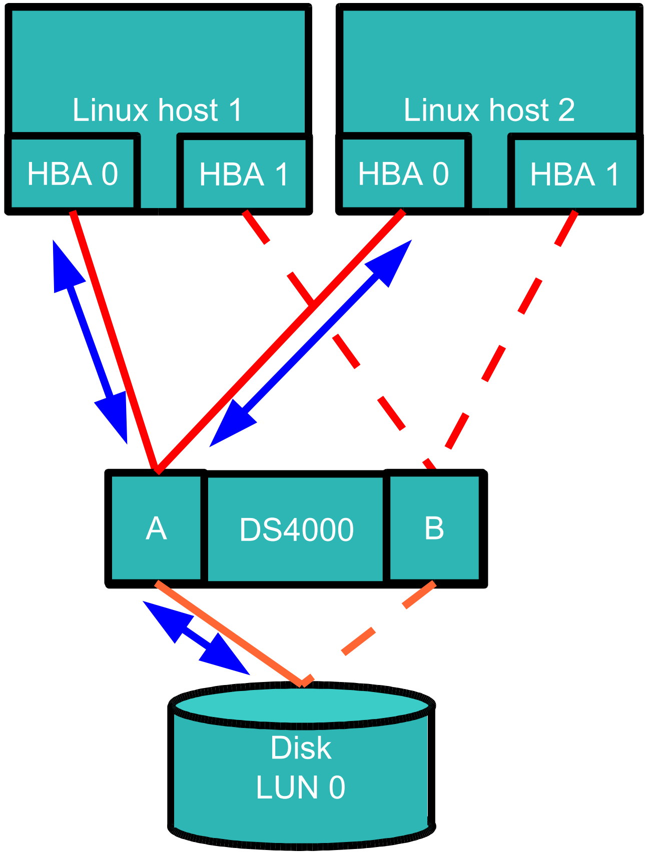




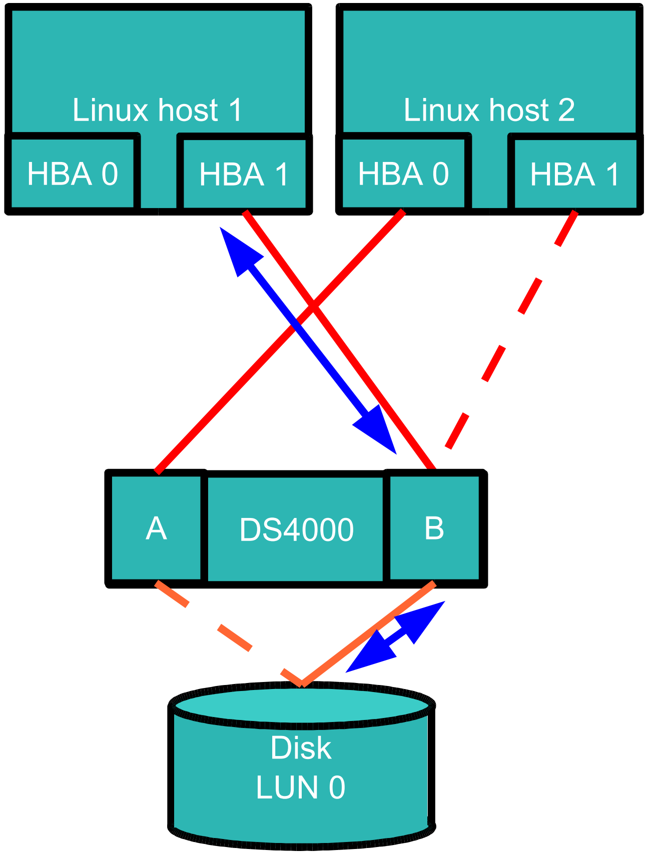
LUN Transfer to Alternate Controller

- DS4000 transfers LUNs to alternate controller
  - Volumes are owned by one controller
  - Volumes can be accessed through both controllers
  - Volume ownership is always transferred to the controller that is used for volume access -> transfer time approx 1 s
- Two multipathing solutions available
  - QLogic failover driver uses AVT
    - Difficult to configure
    - Potential LUN thrashing
  - RDAC uses inband communication
    - Self configuring
    - Suitable for data sharing scenarios

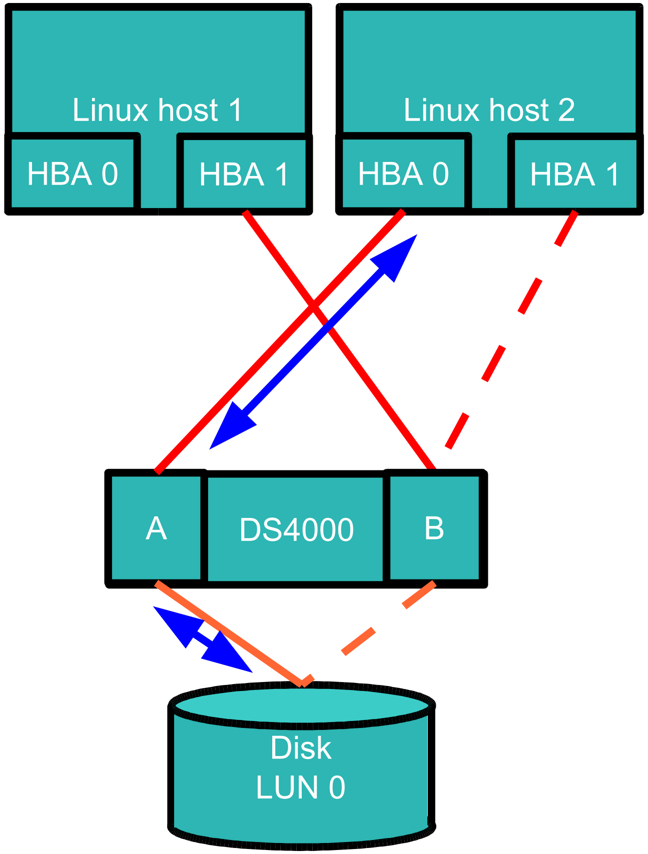
LUN Thrashing Scenario



LUN Thrashing Scenario



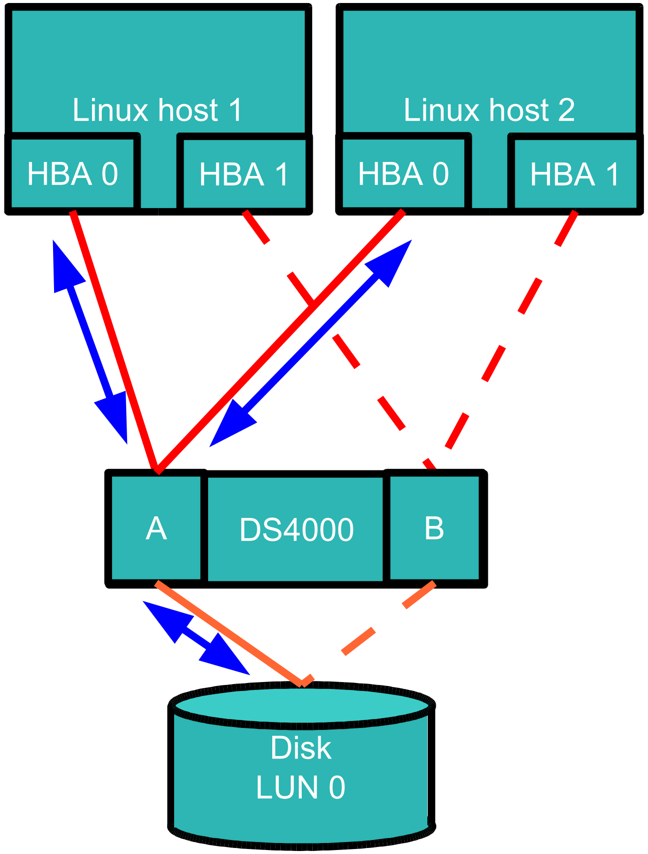
LUN Thrashing Scenario



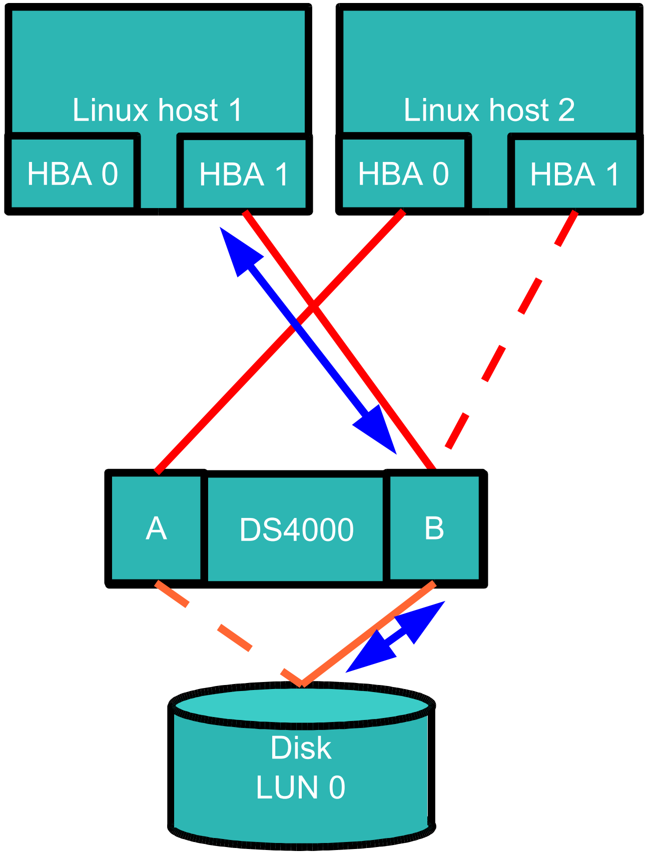
Multipathing with RDAC

- **Must use QLogic non-failover driver**
- **Always uses current path (as reported by DS4000)**
- **RDAC installation**
  - **FC HBA driver must be installed and loaded**
  - **At least one LUN must be assigned and available**
  - **Must use Host Type LNXCLS - AVT turned off**
  - **Must update boot loader configuration**
- **Must run `mppUpdate` after each configuration change**
  - **Updates RDAC configuration files**
  - **Rebuilds Initial RAMDisk**

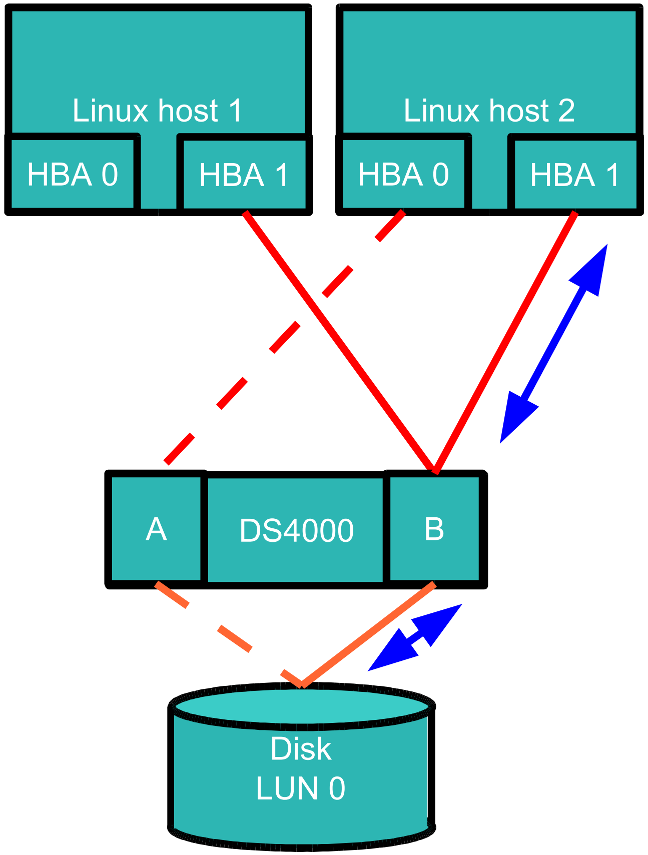
RDAC Shared Data Scenario



RDAC Shared Data Scenario



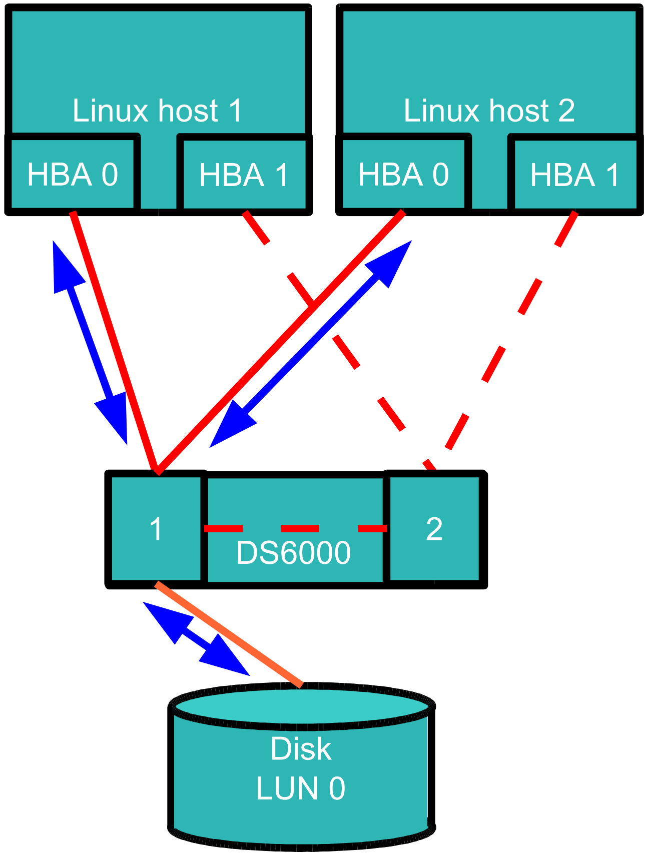
RDAC Shared Data Scenario



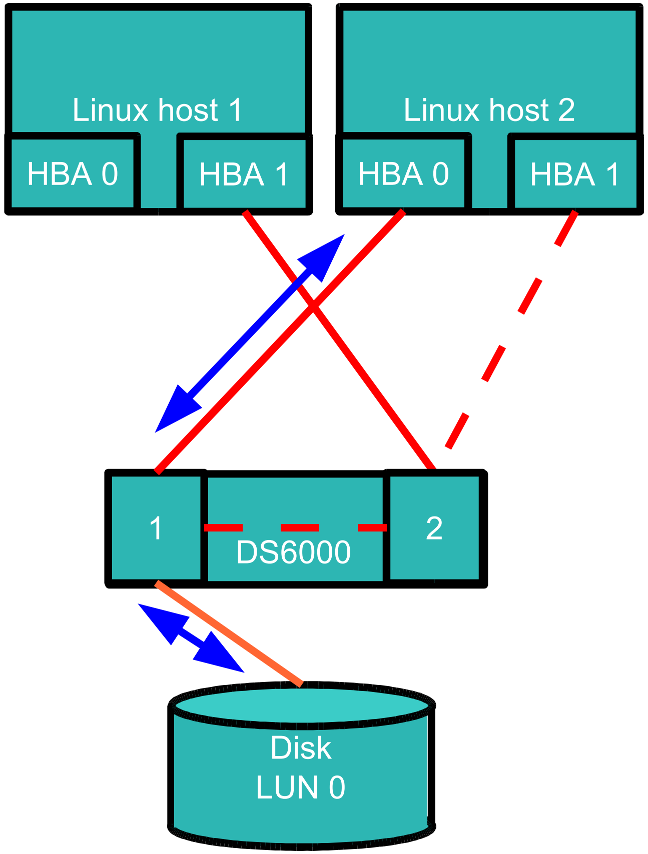
Preferred Path

- DS6000 uses concept of preferred path
  - Volumes are owned by one controller
  - Volumes can be accessed through both controllers
  - Data is transferred to and from owning controller to requesting controller internally -> performance penalty
- SDD knows preferred path automatically
  - Access only through owning controller if possible
  - Dynamic load balancing across ports of preferred controller
- Other multipathing solutions theoretically possible, but must (still) be configured manually

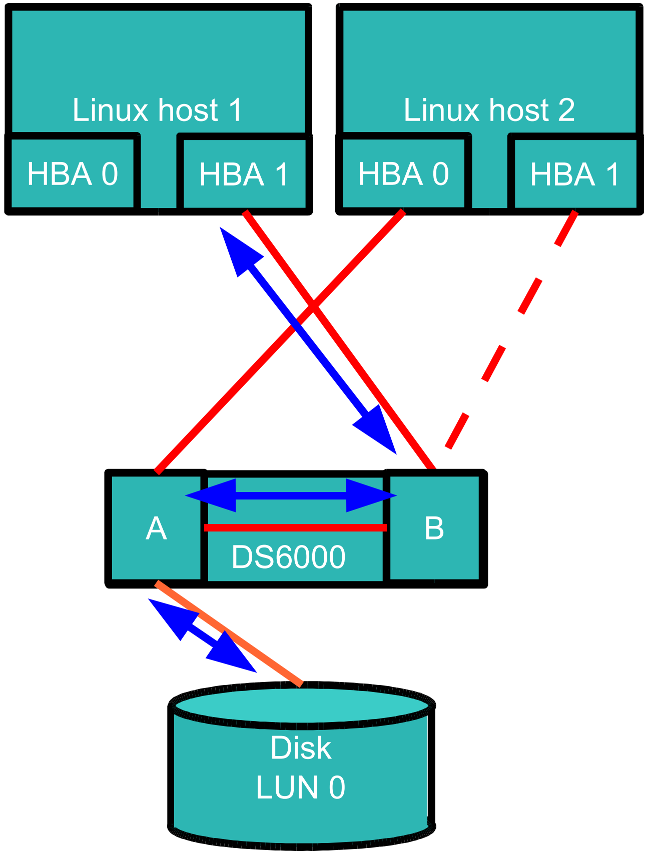
Preferred Path Shared Data Scenario



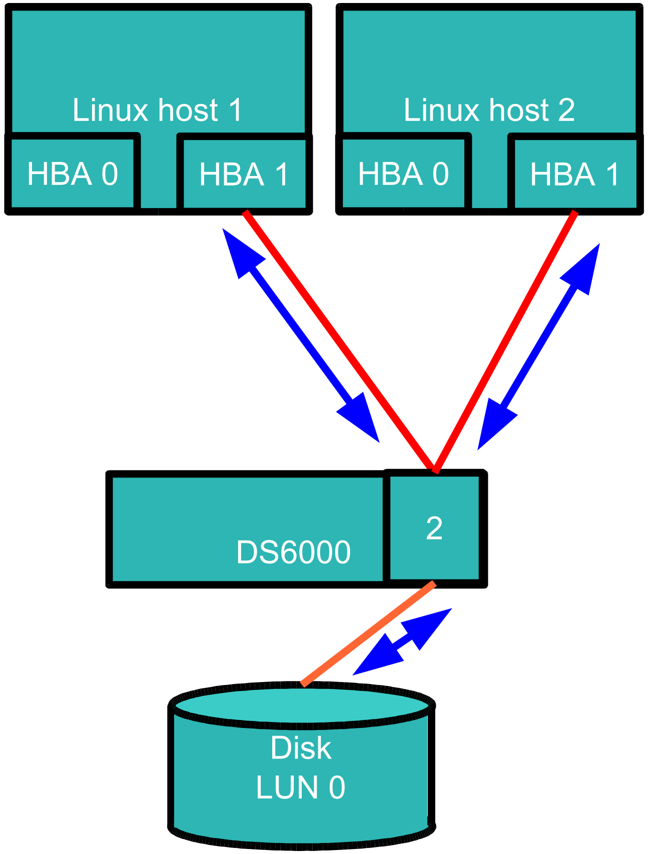
Preferred Path Shared Data Scenario



Preferred Path Shared Data Scenario



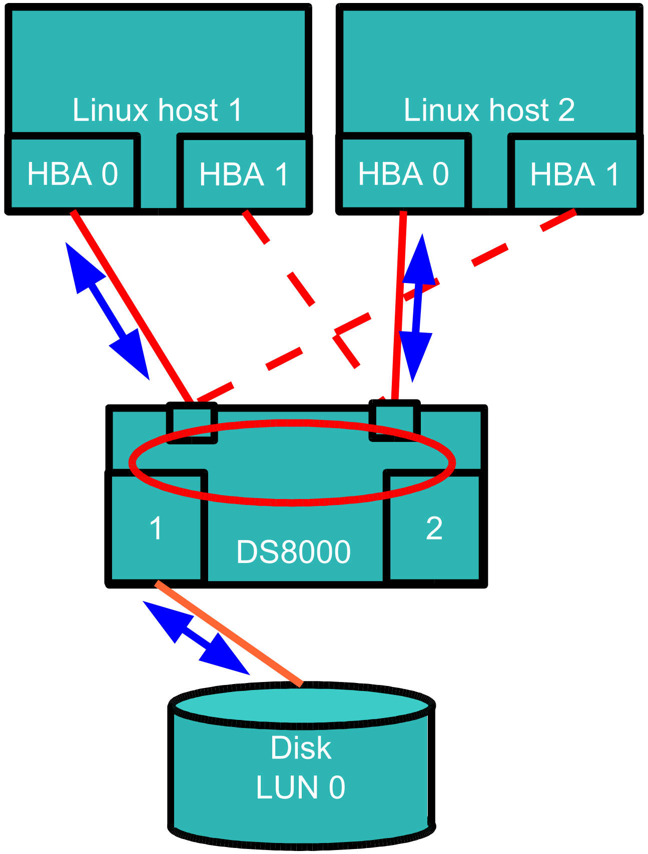
Preferred Path Shared Data Scenario



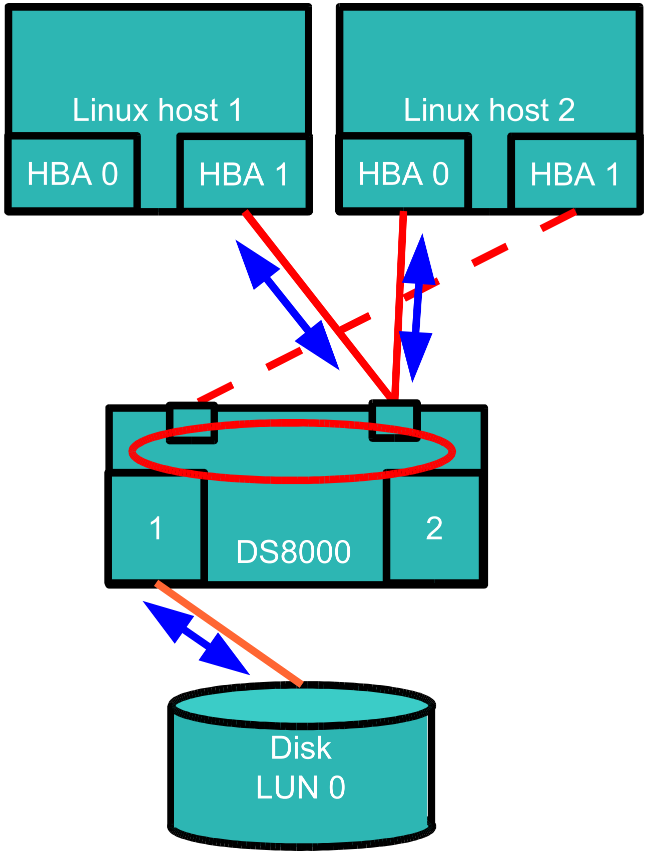
Host Ports Independent of Controller

- **ESS and DS8000 have independent host ports**
  - **Volumes are owned by one controller**
  - **All host ports can communicate with both controllers**
  - **Dynamic load balancing across all ports possible**

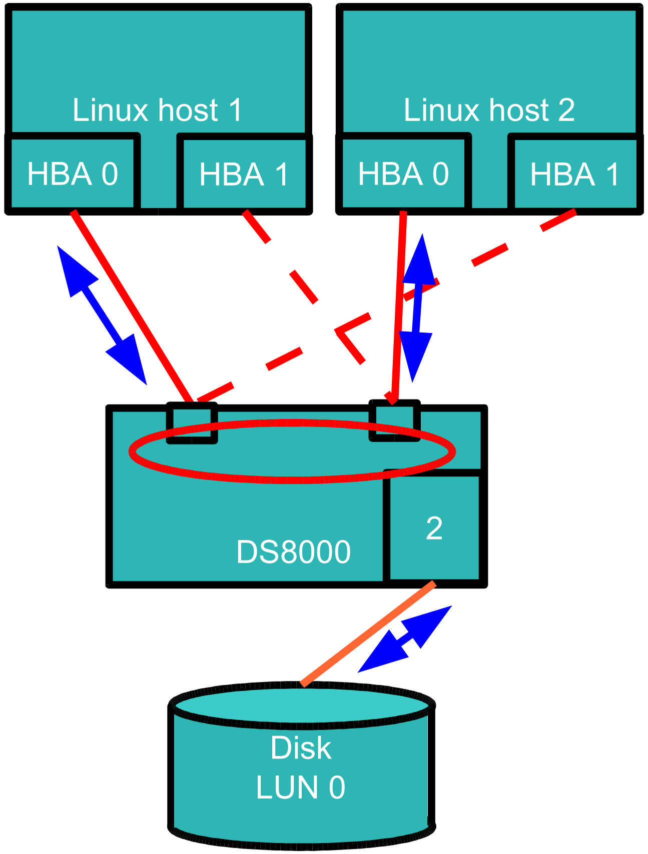
Independent Host Port Shared Data Scenario



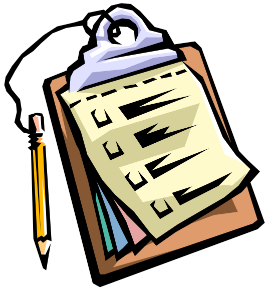
Independent Host Port Shared Data Scenario



Independent Host Port Shared Data Scenario



Tape Specifics



- What is supported
- The Linux SCSI subsystem
- Linux Kernel version 2.6
- Multipathing scenarios
- **Tape specifics**

IBMtape driver

- **For download as binary rpm package**
- **Kernel module `IBMtape.o`**
  - **Required to utilize all LTO capabilies**
  - **Manages medium changer failover**
  - **Provides new devices and ioctl**
- **Daemon `IBMtaped`**

```
NDMC-7:/ # ls -l /dev/IBM*
crw-rw-rw-    1 root     root     253, 128 Sep 25 11:18 /dev/IBMchanger0
crw-r--r--    1 root     root     253, 255 Dec  9 09:41 /dev/IBMtape
crw-rw-rw-    1 root     root     253,   0 Sep 25 11:18 /dev/IBMtape0
crw-rw-rw-    1 root     root     253,  64 Sep 25 11:18 /dev/IBMtape0n
crw-rw-rw-    1 root     root     253,   1 Sep 25 11:18 /dev/IBMtape1
crw-rw-rw-    1 root     root     253,  65 Sep 25 11:18 /dev/IBMtape1n
```

IBMtapeUtil

- **For download as source code**
  - **Exerciser tool**
  - **Software example**
- **Build and install using make**
- **Provides**
  - **IBMtapeutil**
  - **IBMtapeconfig**

```
-------------------------- General Commands: -----------------------------
    1. Open a Device                      7. Request Sense
    2. Close a Device                     8. Log Sense Page
    3. Inquiry                            9. Mode Sense Page
    4. Test Unit Ready                   10. Switch Tape/Changer Device
    5. Reserve Device                    11. Create Special Files
    6. Release Device                    12. Query Driver Version
    Q. Quit IBMtapeutil
------------------------- Medium Changer Commands: -----------------------
   60. Element Information               65. Load/Unload Medium
   61. Position To Element               66. Initialize Element Status
   62. Element Inventory                 67. Prevent/Allow Medium Removal
   63. Exchange Medium                   68. Initialize Element Status Range
   64. Move Medium                       69. Read Device Identifiers
-------------------------- Service Aid Commands: -------------------------
   70. Dump Device                       72. Load Ucode
   71. Force Dump                        73. Reset Drive
--------------------------------------------------------------------------
   99. Back To Main Menu
```

Use LTO Devices

- Native
  - tools: mt, mtx, IBMtapeutil
  - applications: cpio, tar, taper, afio

- 3rd party applications
  - All major backup solutions available for Linux
  - Attention: some are only tested with parallel SCSI attachment
  - Check ISV Martrix for LTO

Native Library Management

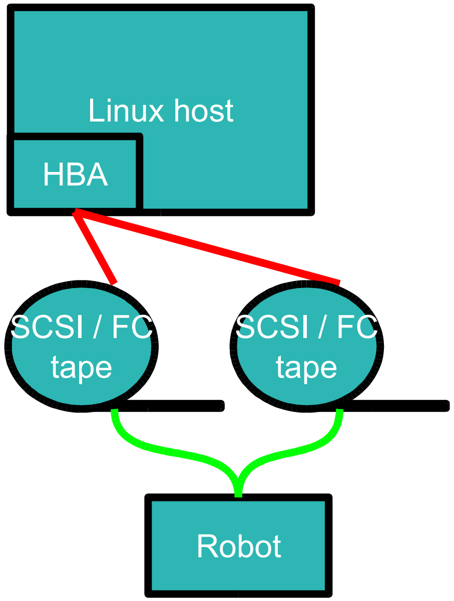
- **Linux tool for media changers: mtx**
- **Media changer is addressed through SCSI generic device**

/dev/sg0 - internal SCSI disk, not relevant here

/dev/sg1 - 1st SCSI tape drive

/dev/sg2 - tape robot (media changer)

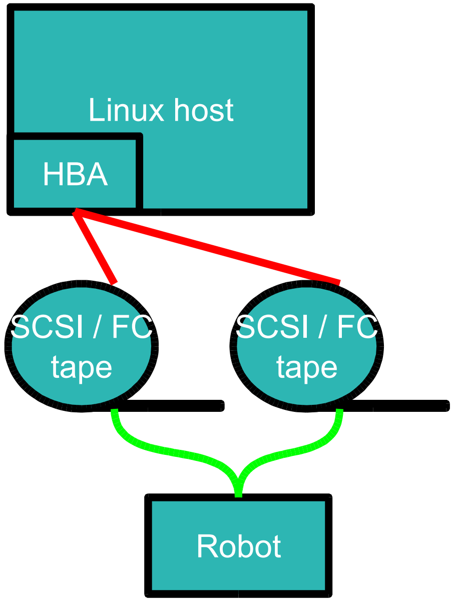
/dev/sg3 – 2nd SCSI tape drive



```
mtx -f /dev/sg1 inquiry
mtx -f /dev/sg2 status
mtx -f /dev/sg2 load <slotnum> [ <drivenum> ]
```

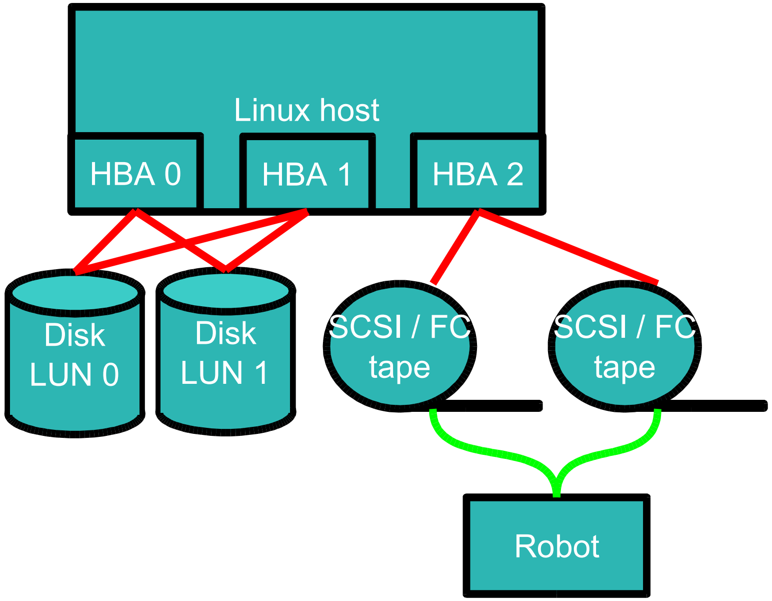
Medium Changer Failover

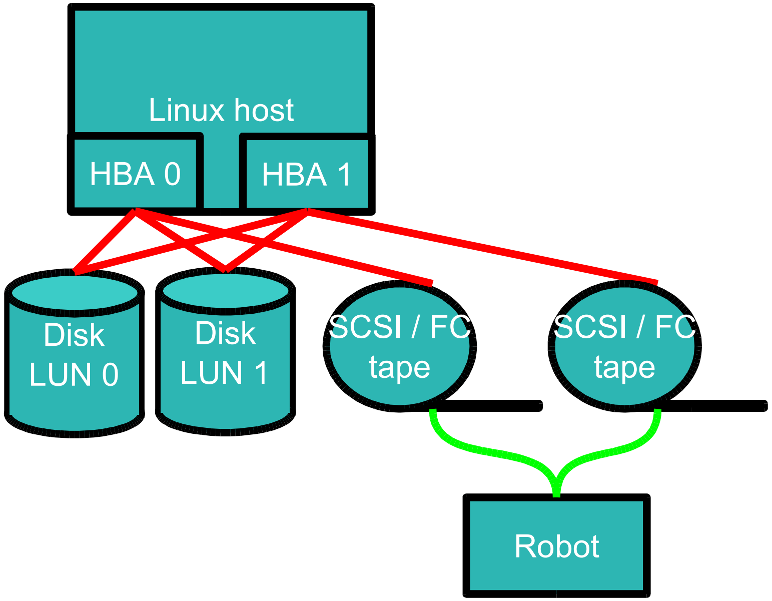
- **Automatitcally moves robot control to another drive in case of a failure**
- **Available for 2582, 3583, 3584**
- **Enabled as an option for IBMtape driver**
- **Check the `/proc/scsi/IBMchanger` file**



Tape and Disk Connected to the Same HBA

- **Possible, but not recommended**
- **Use separate switch zone, too**
- **One driver for all HBAs!**





Questions & Discussion

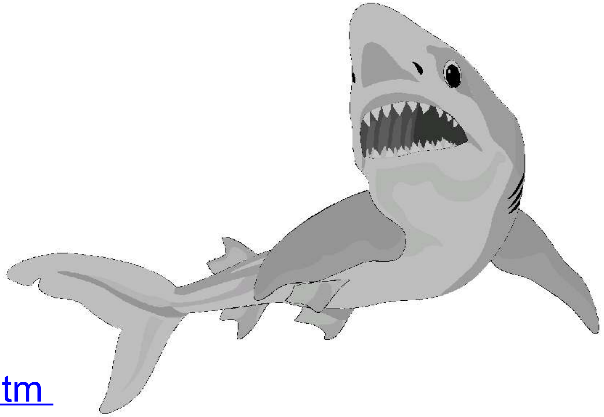
More Questions?

What are your customers needs?

Contact: warmuth@de.ibm.com

ESS / DS6000 / DS8000 Resources

- **Enterprise Storage Server interoperability matrix**
- **Subsystem Device Driver (SDD)**
- **Fibre channel host bus adapter firmware and driver level**
- **Additional supported configurations**
- **ESS host systems attachment guide**



http://www.storage.ibm.com/disk/ess/ess800/supserver.htm

http://www.storage.ibm.com/disk/ds8000/supserver.htm

http://www.storage.ibm.com/disk/ds6000/supserver.htm

DS4000 Resources

- **DS4000 Storage interoperability matrix**
- **Fibre channel host bus adapter firmware and driver level**
- **Additional supported configurations**

http://www.ibm.com/servers/storage/disk/ds4000/interop-matrix.html

- **DS4000 Technical Support**
- **DS4000 Downloads**

http://www.ibm.com/servers/storage/support/disk/

LTO Resources

- **LTO Compatibility Information**
- **LTO ISV Matrix**

http://www.storage.ibm.com/tape/lto/compatibility.html

- **LTO Downloads**

http://www.ibm.com/servers/storage/support/lto/ltodownloads.html

ftp://ftp.software.ibm.com/storage/devdrvr/Linux/

Redbooks

- **Implementing Linux with IBM Disk Storage**
  http://www.redbooks.ibm.com/redbooks/pdfs/sg246261.pdf
- **Linux with xSeries and FAStT: Essentials**
  http://www.redbooks.ibm.com/redbooks/pdfs/sg247026.pdf
- **Implementing IBM LTO in Linux and Windows**
  http://www.redbooks.ibm.com/redbooks/pdfs/sg246268.pdf
- **Linux Clustering with CSM and GPFS**
  http://www.redbooks.ibm.com/redbooks/pdfs/sg246601.pdf

White Papers

- **FAStT and Linux HowTo**

  http://www.ibm.com/developerworks/eserver/articles/install_fibre/index.html

- **FAStT and RH AS Cluster**

  http://www.ibm.com/servers/esdd/articles/redhat/index.html

- **ESS Attachment to United Linux 1 (IA-32)**

  http://www.ibm.com/support/docview.wss?uid=tss1td101235

  http://w3.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD101235

- **Addendum to the Solution Assurance Process**

  http://ulrich.walter.de.userv.ibm.com/portal.htm

Legal Notices

Both Linux and Storage are rapidly changing environments.

This information is presented "as is" without any warranty of any kind. Customers are responsible for determining the suitability to their respective environments.

Only a representative subset of the IBM offerings are presented here. Products not mentioned should not be interpreted as a lack or withdrawal of support of those products.

Information concerning non-IBM products was obtained from a supplier of these products, published announcement material, or other publicly available sources. Questions on the capability of non-IBM products should be addressed to the supplier of those products.

Some information in this presentation addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in official IBM product announcements. All statements regarding IBM future direction and intent are subject to change or withdraw without notice, and represent goals and objectives only.

The information is presented here to communicate IBM's current investment and development activities as a good faith effort to help with our customers' future planning.   Contact your local IBM business contact for details on specific products, programs or services.

The following are trademarks or registered trademarks of the International Business Machines Corporation:

AIX, AS/400, AS/400e, CICS, DB2, DB2 Universal Database, e-business (logo), Enterprise Storage Server, the eServer logo, ESCON, FlashCopy, IBM, Intellistation, iSeries, Magstar, Modular Storage Server, MQSeries, Netfinity, NUMA-Q, OS/390, OS/400, Parallel Sysplex, pSeries, RS/6000, S/390, SANergy, Seascape, Sequent, Sequent (logo), SP, SP2, SSA, StorWatch, Thinkpad, Tivoli, Tivoli Storage Manager, Ultrastar, WebSphere, xSeries, zSeries.

Microsoft, Windows, Windows NT and the Windows logo are registered trademarks of Microsoft Corporation. Intel and Pentium are registered trademarks of Intel Corporation. UNIX is a registered trademark licensed exclusively through the OPEN group. LINUX is a registered trademark of Linus Torvalds. Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc.

Red Hat, the Red Hat "Shadow Man" logo, RPM and all Red Hat-related logos are trademarks or registered trademarks of Red Hat, Inc. Caldera Systems, the C-logo, SCO, and related logo, are trademarks or registered trademarks of Caldera Systems, Inc. Turbolinux and "lightning bolt" logo are registered trademarks of Turbolinux, Inc. SuSE, and SuSE "lizard" logo, are trademarks of SuSE, Inc.

Linux is a registered trademark of Linus Torvalds

Other company, product and service names may be trademarks or service marks of others.